Social-Topical Affiliations: The Interplay between Structure and Popularity

Daniel M. Romero*
Center for Applied Math
Cornell University
Ithaca, NY
dmr239@cornell.edu

Chenhao Tan*
Dept. of Computer Science
Cornell University
Ithaca, NY
chenhao@cs.cornell.edu

Johan Ugander*
Center for Applied Math
Cornell University
Ithaca, NY
jhu5@cornell.edu

ABSTRACT

Information popularity and social relationships are intimately connected. However, measuring the extent to which they affect each other has remained an open question. Because we now have access to rich and large data sets from online social networks, we can begin to quantitatively understand the interplay between them.

We examine the interface of two decisive structures forming the backbone of online social media: the graph structure of social networks — who is friends with whom — and the set structure of topical affiliations — who talks about what. In studying this interface, we identify key relationships whereby each of these structures can be understood in terms of the other. The context for our study is Twitter, where we look at the social network of both follower relationships and communication relationships, alongside the affiliations outlined by the hashtags used by people to label their communications.

On Twitter, we demonstrate how the hashtags that a user adopts can be used to predict their social relationships, and also how the social relationships between the adopters of a hashtag can be used to predict the future popularity of that hashtag. Importantly, we find that both relationships are driven by highly computationally simple structural determinants. While our analysis focuses on Twitter, we view our analysis of social-topical affiliations as broadly applicable to a host of diverse affiliations, including the movies people watch, the brands people like, or the locations people frequent.

Categories and Subject Descriptors

H.1.2 [**Information Systems**]: Models and Principles—*Human information processing*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*; H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces—*Collaborative computing, Web-based interaction*

General Terms

Algorithms, Experiments

*All authors contributed equally to this work.

Keywords

graph structures, link prediction, popularity prediction, social networks, socio-topical affiliations, social media

1. INTRODUCTION

Online social networks and online information sharing has gained tremendous popularity over the passed several years. This has provided researchers with the opportunity to track how relationships form and how information diffuses online. There have been many studies that investigate how edges in social networks form and how networks evolve [4, 22, 11, 15, 18]. Furthermore, many studies have looked at online information sharing, either on social tag systems [7, 10, 21, 24, 8] or on the mechanisms of information diffusion

[28, 6, 12].

It is an interesting question whether the information held by individuals in the network could describe properties of the existing social network, and whether the social network itself could describe the properties of the information that diffuses on it. We could think of a system where each user in a network is tagged with a certain topic if information related to that topic passes through the user, or if there is evidence that the user is interested in the topic. For example, the user may have downloaded a movie about the topic. This would generate a topical affiliation system that could be useful in understanding the structure and evolution of the network itself. In this work, we aim to bridge both the social and informational aspects together and study the extent to which they are related and can predict each other. To do so, we look at the intersection of two key structures of online social media – the set structure of topical affiliations and the graph structure of social networks. We aim to understand the interplay between the two, thereby understanding "the social structure of topics", and the "topical structure of social."

We begin by looking at the classical problem of link prediction in social networks. Many people have studied this problem and the approach has often been to look at the features of the existing network in order to predict future connections [19, 30, 29]. In this work, we use features related to the topical proximity of the users as well as the graph properties of these topics, and demonstrate how prediction models based on topical features can yield impressive prediction accuracy. This part of the paper shows how the topical structure of the users in the network can inform our understanding of the structure of the network itself.

Next, we develop our main result, understanding the extent to which the structure of the network determines the topical information diffusion process. In particular, we are interested in whether the structure of the graph induced by the initial set of adopters of a certain topic can tell us something about the eventual popularity of

the topic. Here, we are using "topic" in a loose sense, referring to a product, idea, or even a behavior. The idea that the speed and magnitude of adoption of products, ideas, and behaviors can be driven by "viral marketing" techniques has gained tremendous popularity over the past few years [2, 5, 25, 17, 26]. The premise is that one can utilize the edges of an existing social network as bridges for information to spread from person to person. A common question is what kind of topics will go viral in the future, and the focus of our study is to ask: how precisely is this related to graph structure? To shed light into this question, we test a predictive algorithm that takes as features properties of the induced subgraph of the early adopters of a topic, with the goal of predicting its eventual popularity.

We find that the structure of the early adopter graphs can indeed have predictive power about the popularity of the topic. Furthermore, we observe that the relationship between the topological properties of the initial graphs and the popularity of the topic is not always as expected. For example, popularity of a topic does not monotonically change with the number of social connections among the initial users. Instead, we find that the topic exhibits high future popularity when the number of connections is either very high or very low. This could come as a surprise since few connections among adopters of a topic could be perceived as the topic lacking "virality."

Our particular domain of study is Twitter. We use *hashtags* – labels that users include in their posts to indicate the topic of the message – to distinguish between different topics and the follower and @-message network as a proxy for social connections among the users. Because these networks are directed and the @-message network is weighted, we are able to compare the differences in our results when considering reciprocate and unreciprocated relationships, as well as strong and weak ties. We find that while strong and mutual ties are easier to predict, they are less useful than weaker directed ties when predicting hashtag popularity.

Paper organization The organization of our paper is as follows: Section 2 presents the details of the data set forming the basis of our study. Section 3 studies the prediction of directed ties, in various forms, from the common hashtag usage of user pairs. Section 4 investigates how the social structures of hashtags early adopters can be used to predict the future popularity of those hashtags. Section 5 summarizes related work. Finally, Section 6 concludes.

2. DATASET

The dataset used in this paper consists of two main parts: hashtags and networks. From August 2009 until January 2010 Twitter was crawled using their publicly available API. The last 3,200 tweets of each existing user were collected. Overall, over three billion messages from more than 60 million users were obtained during this crawl [28].

Hashtags. A convention widely used by Twitter users is a tagging system where a user includes a single undelimited string proceeded by a "#" character. This string is referred as *hashtag* and it is meant to label the tweet so other people know what it is about, the designate that it belongs to a particular conversation topic. For example: "*What a game last night between the Thunder and Grizzlies. Was up till 1am watching that triple over-time thriller. #NBA.*" We extract all the hashtags that have appeared in our dataset and users who have utilized at least one hashtag. Our data set contains a total of 7,305,414 hashtags and 5,513,587 users who utilized at least one hashtag. On average, each hashtag is used by 9.48 distinct users, while a user posted about 12.57 different hashtags.

Graphs. We get the follower/followee network from [16], which contained the list of people each person was following at crawl time. If user A follows user B we create the edge (A, B). There are around 366 million edges among the users who have utilized at least one hashtag. The second graph is based on *@-messages*. These are posts that are publicly directed from one user to the other. They can be used as directed messages, or to reference another user. To send an @-message, the sending user includes the "@" character next to the receiving person's user name in their tweet. The @-graph is created by constructing the edge (A, B) if A has sent at least n @-messages to B in the tweets available (n is a threshold to indicate the strength of the relationship between two users). There are around 85 million edges in the @-graph with threshold $n = 1$ among the users who have utilized at least one hashtag.

3. LINK PREDICTION

In this section, we ask to what extent the hashtags that an individual has used reveals their ties to other users in Twitter's directed social graph, or their ties to other users via @-communication. By allowing hashtags to define user sets, we can view Twitter users as embedded in the set system of these hashtags. We begin by characterizing the features of the hashtag usage of two individuals that we wish to process. We then consider a prediction problem, where we are trying to predict the presence of an edge between arbitrary pairs of individuals. From this, we observe that the size of the smallest common hashtag that two users overlap on is a surprisingly informative predictor. Having observed this, we ask to what extent the graph structure of these smallest common hashtags can be used to improve prediction accuracy, ultimately obtaining remarkably a capable predictive model.

3.1 Measuring hashtag distance

In order to approach this question, we must first summarize the hashtag usage similarity of two individuals into features that could plausibly serve as similarity/distance measures. Perhaps the most obvious measure of similarity is the number of hashtags that two users have in common. This measure is immediately problematic, since it does not distinguish between hashtags that are broadly adopted and those that have only been used by a handful of users. More appropriately, we consider features that relate to the frequency of the common hashtags in the broader population. For this, we consider the size of the smallest common hashtag, the size of the largest, the average size, and also two measures that aggregate the common overlap of the full sets: the sum distance and the Adamic-Adar distance [1].

Consider the following notation:

- Let $u_1, \ldots, u_N$ be the N users.
- Let $h_1, \ldots, h_M$ be the M hashtags.
- Let $H(u_i)$, be the set of hashtags used by users u_i.
- Let $U(h_j)$ be the users who used hashtag h_j.

This is the structural information we aim to process. The features of the common hashtags between users that we consider in the work are:

- The number of hashtags in common, $|H(u_i) \cap H(u_j)|$.
- The size of the smallest common hashtag, $\min_{h \in H(u) \cap H(v)} |U(h)|$.
- The size of the largest common hashtag, $\max_{h \in H(u) \cap H(v)} |U(h)|$.

- The average size of the common hashtags, $\frac{1}{|H(u)\cap H(v)|}\sum_{h\in H(u)\cap H(v)}|U(h)|$.
- The sum of the inverse sizes, $\sum_{h\in H(u)\cap H(v)} 1/|U(h)|$.
- The Adamic-Adar distance, $\sum_{h\in H(u)\cap H(v)} 1/\log|U(h)|$.

The size of the smallest common hashtag is an intuitively attractive measure: it captures the extent to which the conversations two persons share are unique or not. In [14], Kleinberg studied social networks where individuals were viewed as embedded in a set system, much like our hashtag set system, and an individual u was linked to an individual v with probability proportional to $d(u,v)^{-a}$, where $d(u,v)$ is the size of the smallest common set. Kleinberg showed that decentralized greedy routing with regard to this measure takes polylogarithmic time if and only if $a = 1$. Beyond studying the performance of the minimum common hashtag size in a predictive setting, we therefore also investigate the extent to which such an inverse proportional dependence holds true in our setting.

The size of the largest common hashtag is an intuitively poor feature for predicting links, and we include it here specifically as a control of sorts, showing that not all features of the common hashtags are informative. Very commonly, the largest common hashtag that users overlap on is one of a few extremely popular hashtags, for example #musicmonday, #ff, or #fail.

In choosing to study the sum distance and Adamic-Adar distance — a measure introduced for studying the similarity of web-based social networks derived from common homepage content [1] — we allow ourselves to consider similarity measures using all common sets.

3.2 Predictive model

Given the features defined above, we now investigate how well the topical overlap represented in the common hashtags allow us to predict the presence of links. We formulate this task as a balanced classification task: given a set of 100,000 users that coincide on some hashtag, where 50,000 are disconnected pairs and 50,000 are connected, what sort of prediction accuracy can we obtain, compared to a naive 50% baseline?

Performing classification against completely arbitrary user pairs would have been over generous information, since arbitrary users rarely coincide on any hashtags, and the coincidence rates between connected users is understandably going to be higher. In fact, 35% of user pairs where one user follows the other coincide on some hashtag, and fully 78% coincide when comparing user pairs where one person has @-messaged the other person at least once. But even among completely arbitrary hashtag-using users, coincidence is actually rather common: fully 4.9% of arbitrary user pairs coincide on one or more hashtags.

We approach the problem using logistic regression with 10-fold cross-validation. For all six features, in addition to a linear term, we also include the logarithm and the inverse of each value, allowing us to more robustly extract non-linear dependencies.

We consider the tasks of predicting follow edges, mutual follow edges, @-message edges, and mutual @-message edges.

It is natural to view the number of @-edges as the strength of a tie, and we therefore also consider our classification task applied to @-edges thresholded on high @-message counts. The accuracies we obtain are shown in Table 3.2, where we report the performance of both a full model, considering all features under all transformations, and also models trained on a single feature set (where we include its two transformations).

We see that classification based on common hashtag usage exhibits powerful prediction accuracy: for follow edges our accuracy is 74%, for @-edges it is 83%, and for strong @-edges (more than 20 messages), our accuracy is 87%. We achieve comparable performance when predicting mutual edge relationships. Beyond the impressive performance, we note that the size of the smallest common hashtag is a consistently accurate feature when considered alone, especially when trying to predict strong ties. As expected, the size of the largest common hashtag is least performative.

3.3 Predictive model with edges

The features we consider above do not extract anything about the graph structure of the induced subgraphs that each hashtag defines. Given the accuracy of models based solely on the size of the smallest common hashtag, here we chose to investigate to what extent the social structure of these smallest common hashtags — the most unique topic that two users have in common — can further improve our predictions.

One motivation for considering the graph structure is the observation, shown in Figure 1, that the edge density for similarly sized hashtags can differ considerably. Simply put, some hashtags — some topics — are much more 'social' than others. For social-topical contexts involving geography, this would translate to knowing that two people both visit a bar versus knowing that they both visit the same bank. Both may be equally popular locations to visit, but the bar is a decidedly more social environment, and visiting the same bar is more likely to predict social interaction than visiting the same bank.

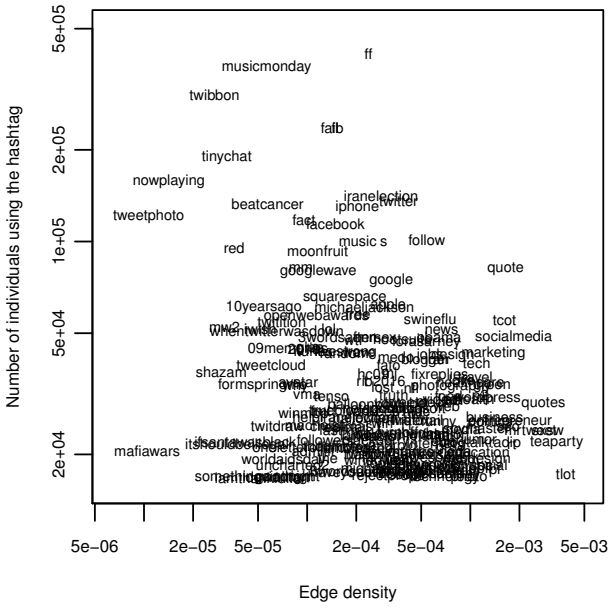

Figure 1: Edge density heterogeneity for the 100 most common hashtags in the dataset.

When performing this classification, it is important to avoid inadvertently incorporating a circular reference whereby the link is directly present in the feature. Consider the problem of predicting an edge between a pair of users where the induced subgraph for their smallest common hashtag has two nodes and one edge. If we don't make this correction, we would know for certain that these two users are connected. We therefore let our edge count feature

Directed edges

Model Features	Follow	@ ≥ 1	@ ≥ 3	@ ≥ 5	@ ≥ 7	@ ≥ 9	@ ≥ 20
All hashtag features	0.737	0.826	0.850	0.860	0.862	0.870	0.871
# common HTs	0.713	0.781	0.798	0.800	0.804	0.809	0.816
Smallest HT size	0.703	0.799	0.828	0.841	0.842	0.854	0.855
Largest HT size	0.582	0.587	0.584	0.585	0.581	0.583	0.575
Average HT size	0.589	0.662	0.683	0.702	0.697	0.723	0.720
Sum distance	0.712	0.804	0.832	0.845	0.848	0.858	0.860
Adamic-Adar distance	0.727	0.809	0.831	0.842	0.846	0.852	0.856
Hashtag features + Edges	**0.766**	**0.863**	**0.889**	**0.921**	**0.940**	**0.949**	**0.976**
Edges of smallest	0.647	0.790	0.816	0.827	0.865	0.872	0.886

Mutual edges

Model Features	Follow	@ ≥ 1	@ ≥ 3	@ ≥ 5	@ ≥ 7	@ ≥ 9	@ ≥ 20
All hashtag features	0.762	0.827	0.868	0.869	0.868	0.867	0.866
# common HTs	0.739	0.782	0.809	0.813	0.812	0.812	0.808
Smallest HT size	0.715	0.803	0.849	0.853	0.852	0.852	0.856
Largest HT size	0.576	0.562	0.590	0.583	0.574	0.569	0.548
Average HT size	0.597	0.671	0.712	0.706	0.707	0.706	0.743
Sum distance	0.725	0.808	0.854	0.857	0.856	0.856	0.860
Adamic-Adar distance	0.751	0.807	0.850	0.854	0.852	0.852	0.849
Hashtag features + Edges	**0.796**	**0.864**	**0.922**	**0.936**	**0.934**	**0.949**	**0.967**
Edges of smallest	0.651	0.788	0.829	0.832	0.833	0.837	0.861

Table 1: Prediction accuracies for directed and mutual edges, as trained on the full set of hashtag features, trained on individual hashtag features, and also trained on edge features. Accuracy was evaluated using 10-fold cross-validation on a balanced classification dataset.

encode the number of edges present in the smallest common hashtag between users *other* than the two users being considered.

By including the count of such edges appearing in the smallest common hashtag as a feature, where the type of the edges is the same as the type we are trying to predict, we see that our classification performance becomes remarkably accurate, demonstrating an accuracy of 76.6% when classifying follower relationships and 97.6% when classifying strongly tied @-edges.

3.4 Routing

In this section we briefly discuss whether the user-generated hashtag set system is amenable to greedy routing according to the smallest common set distance measure. From Figure 2, we see that aside from considerable heterogeneity, the probability of a link as a function of minimum common hashtag distance very much appears to obey an inverse power law. What In the case of @-communication (here thresholded on 3 messages), the probability of linkage appears to be moderately close to the $a = 1$ necessary for efficient decentralized routing.

What would greedy routing via hashtags on twitter mean? In practice, this would mean that if user u was trying to route a message to user v via the Twitter social graph, using only knowledge of what hashtags user v had used, they would greedily pass the message to their graph neighbor who was closest to v in the above defined 'minimal common hashtag' distance, and instruct that neighbor to pass the message along using the same greedy heuristic. If the social graph were perfectly embedded in the set system with the required structure, specifically with $a = 1$, then the number of steps needed to route the message would be only polylogarithmic in the number of Twitter users, which should be considered surprisingly efficient.

For the purposes of routing information on Twitter, this procedure would not be of practical interest, but observing the presence of this structure does however have serious implications for understanding social networks in a much broader sense. To the extent that follow and @-communication behavior on Twitter reflects social structure in society at large, observing this structure becomes a statement about how to find people in society based only on interests: if you are looking to meet with a particular person, and all you know are their 'interests' in some general sense, this result dictates that you should be able to efficiently approach them through your social network by navigating your search greedily with regard to these interests. Previous studies of online social networks have suggested that greedy routing with regard to geographic distance is very nearly successful in this sense [23], but to our knowledge this is the first study to empirically investigate greedy routing on social networks purely based on interests.

Computing these link probabilities required evaluating the distances of both the linked users as well as the distances of all non-linked pairs of users, a significant hurdle. Because there are over 5 million users in our Twitter dataset, this computation is not practically feasible. Instead, our methodology for circumventing this problem was to sampled 10^9 pairs of users uniformly at random, with replacement. We then compared the number of edges spanning a given distance to the estimated number of total user pairs for that same distance, given the sample.

4. SOCIAL ADOPTION OF HASHTAGS AND FUTURE USERS

As we observed in the previous section, the structure of the graph can be important in predicting links. In this section, we aim to exploit the information embedded in the structure of the network by using it to predict the future popularity of the hashtags. In particular, we study the extent to which the connections among the initial set of adopters of a hashtag can predict its eventual popularity.

4.1 Correlations between initial graph structure and popularity

The properties of the graph of initial adopters of a hashtag tell us about the social structure of the initial adopters. This can in turn

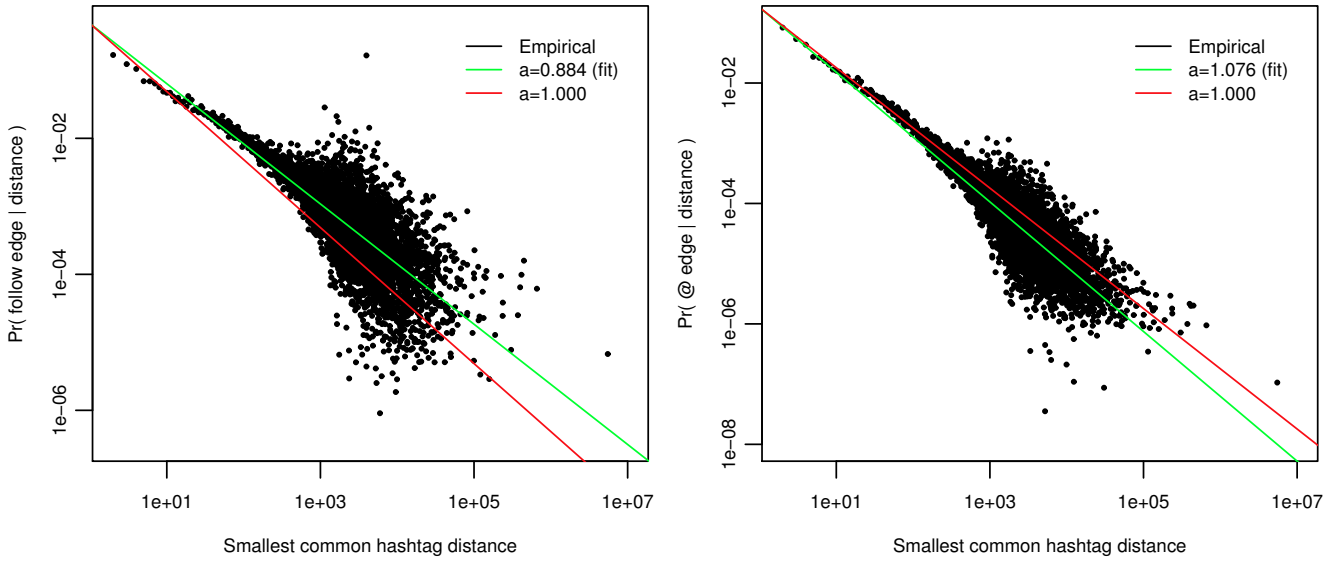

Figure 2: Linkage probability as a function of smallest common hashtag. (a) The probability of a given user following another user as a function of the size, and (b) the probability of a given user @-messaging another user as a function of size. Both figures are log-log scale.

suggest properties of the diffusion mechanism of the hashtag. For example, if the graph has a very large number of edges in it, it suggests that the users could have found out about the hashtag from each other and that many of the friends who haven't adopted it yet are likely to do so in the future. On the other hand, if there are very few edges in the graph, this suggests that the initial adopters did not discover the hashtag through their connections, since their connections had not adopted it yet, which means that users are not "virally" adopting the hashtag. It is an interesting question whether the eventual growth of the hashtag depends on the number of edges in the initial graph, and if so, whether large growth is correlated with a high or a low number of edges. Of course, more detailed properties of the graphs such as the number of connected components, the number of singletons, and the size of the largest component could also be important.

We begin by exploring how different structural properties of the initial graph affect the probability that a hashtag will grow. We consider all 7397 hashtags in our data that had at least 1000 adopters, and construct the follower graph induced by these 1000 users. For each hashtag, we look at the number of users that eventually used the hashtag and compute the number of edges and singletons in their corresponding initial graphs. Figures 3(b) and 3(a) show the that number of eventual adopters does not monotonically change with the number of singletons or edges, instead we find an interior minimum. This suggests that hashtags with either many or few edges and singletons tend to grow more than hashtags with an intermediate number of singletons and edges.

In practice, often times one is not interested in the exact final number of adopters in a diffusion process, instead it is desirable to know if the number of adopters will surpass a certain threshold or if the adopter population will double, triple, etc. In Figures 3(d) and 3(c) we plot the probability that a hashtag with 1000 adopters will reach 1500, 1750, 2000, 2500, 3500, and 4000 adopters as a function of the number of singletons and edges in the subgraph of the initial 1000 adopters. As we found when we were asking about the final number of adopters, the likelihood of growth is highest when the singleton and edge counts are either very large or very small. Furthermore, the trends are consistent for the different choices of k, suggesting that the trend holds for the short, medium, and long terms. Note that we conducted the same experiment with different numbers of initial adopters (the 2000 initial adopters and the 4000 initial adopters), and we observe the same results.

4.2 Exogenous forces vs. Virality.

We observe that hashtags that exhibit large growth tend to be those that either have initial adopters with a large number of connections among each other, or very few connections among each other. However, those hashtags that have a medium number of connections among initial adopters do not tend to grow as much. While we have not conducted experiments that try to find evidence for any theories that explain this phenomenon, we discuss possible explanations.

Let us first consider why a small number of connections among initial users could imply large growth of the hashtag. Many of the hashtags our data set correspond to very popular world events or topics. For example *#iranelection* was used during the disputed 2009 elections, *#michaeljackson* was used around the time of the death of Michael Jackson, and *#iphone* is used around the time a new version of the iphone is released. If we think about the first set of adopters of this kinds of hashtags, it would not be surprising to encounter very little number of connections among them. The reason is that what these hashtags have in common is an force exogenous to Twitter connections that makes people use them on Twitter independently of their connections . For example, the first set of people that used *#michaeljackson* could be the people that first found out about his death on the news and they were probably not friends on Twitter. Hence, a possible explanation for the fact that hashtags with sparse initial graphs tend to grow is that their initial graphs are sparse because the initial adopters are using the hashtag because of an exogenous force, and that exogenous for will also be responsible for the large growth of the hashtag.

On the other hand, we observed that hashtags with very dense

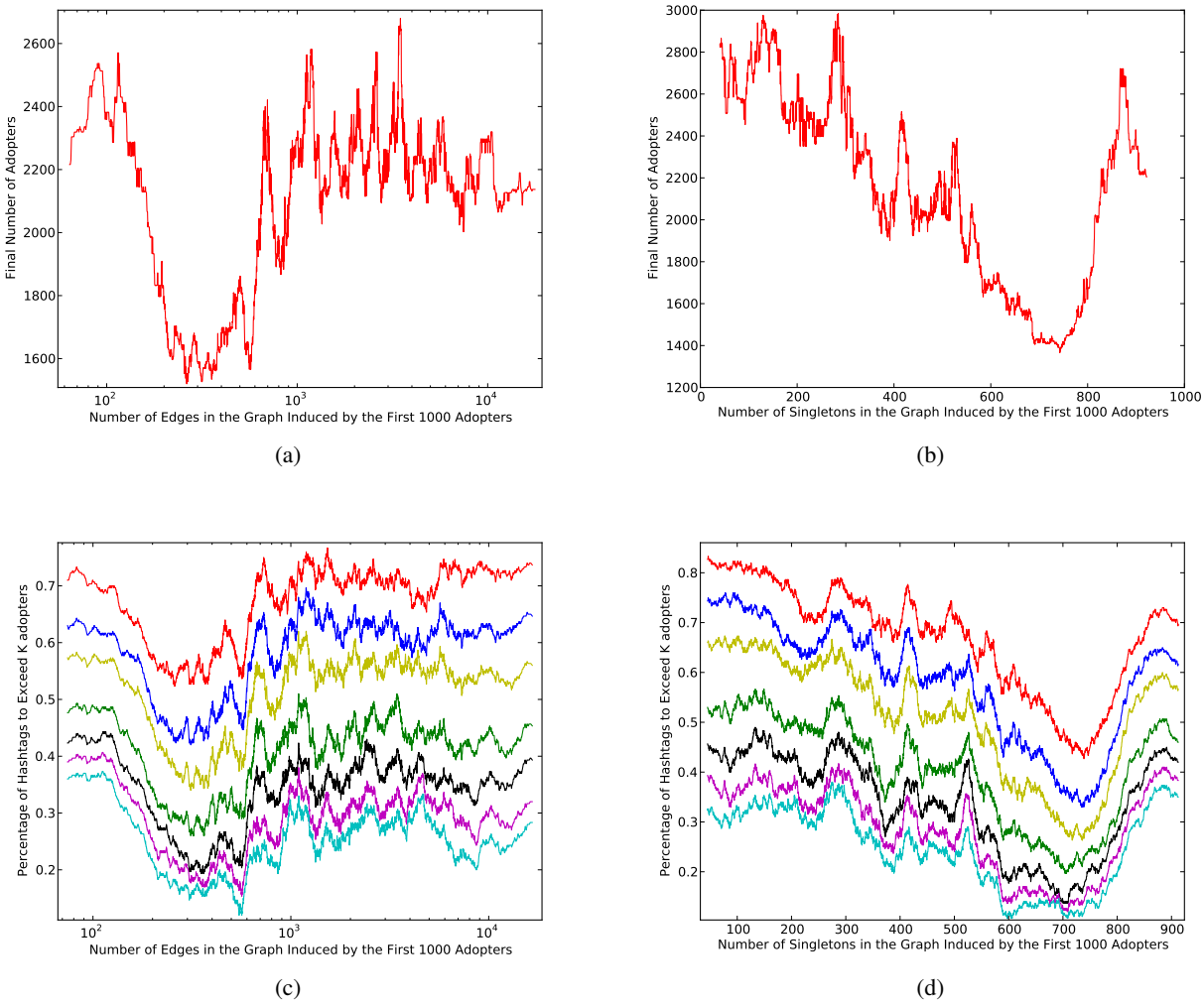

Figure 3: Median number of final adopters as a function of the number of (a) edges and (b) singletons in the graph induced by the 1000 initial adopters, using a sliding window. Probability that hashtags will exceed k adopters given the number of (c) edges and (d) singletons in the graph induced by the 1000 initial adopters, using a sliding window. From top to bottom, $K = 1500, 1750, 2000, 2500, 3000, 3500, 4000$. We observe that hashtags with many or few singletons and edges are more likely to grow than hashtags with intermediate amounts.

initial graphs also tend to grow large. This can be explained by the "virality" argument: if the initial set of users are very well connected among each other, it must be that the hashtag is very sticky. That is, once a users get exposed to the hashtag after seeing that her friend used it, she is very likely to use it. Then her friends that still haven't adopted her will follow her example, and so on. Hence, the dense initial graph of a hashtag may signal that the hashtag is "going viral" and that explains why it will eventually obtain many adopters.

The hashtags that are in the middle of these two extremes lack both virality and exogenous force and hence do not obtain many adopters. Interestingly, we find that these two competing effects generate an interior minimum that we can observe at large scale in our data. However, we note that these theories are only meant as possible explanations and it is an open problem to design experiments that provide further evidence that these are the mechanism that explain the observed interior minimum.

4.3 Predicting growth from structure

We have seen how the number of singletons and edges of the initial graph can be informative with respect to a hashtag's growth. Now we would like to study of these two features, and more generally, the structure of the initial graph can actually predict whether the hashtag will obtain many additional adopters. In order to select appropriate features for a prediction model, it is important to understand the different kinds of connections we can differentiate on Twitter based on the following graph.

Informational and social edges. In Twitter, users can unilaterally follow or @-message other users without having to ask for their approval. This environment allows for the connections of users to have different meanings. For example, if two users on Twitter are friends in real life, they may be likely to follow each other. On the other hand, if a user on Twitter is interested in another users, but they do not actually know each other, we would expect the following relation may be unreciprocated. We refer to this type of

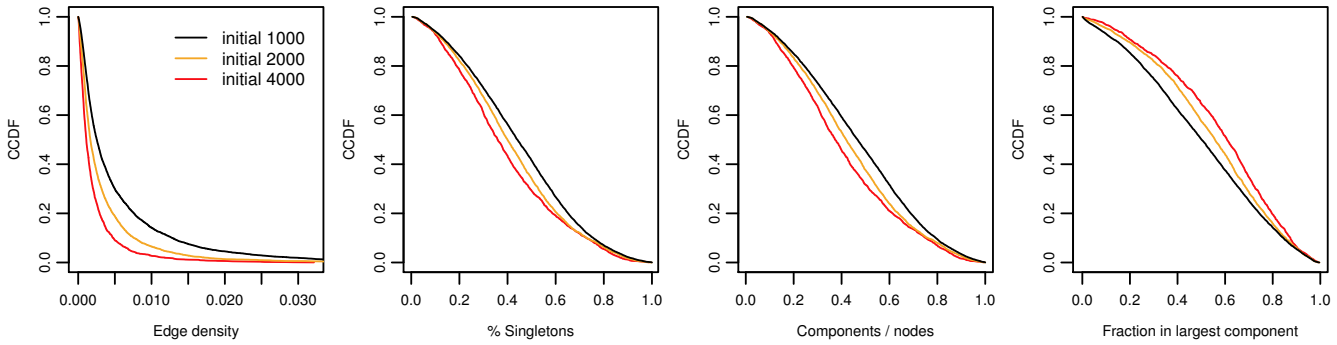

Figure 4: Distribution of the structural features of the subgraphs induced by the 1000, 2000, and 4000 initial adopters. We see that while the edge count exhibits a heavy tailed distribution, the number of singletons, components and the size of the largest component are all broadly distributed over their support.

Model Features	F: 1k→2k	F: 2k→4k	F: 4k → 8k	@≥3: 1k→2k	@≥3: 2k→4k	@≥3: 4k→8k
All	**0.674**	**0.668**	**0.683**	**0.568**	0.574	0.590
Social Graph Only	0.639	0.639	0.642	0.565	0.568	0.590
Full Graph Only	0.658	0.659	0.658	0.568	0.571	0.581
Info. Graph Only	0.642	0.649	0.663	0.559	**0.576**	**0.601**
# Full Graph Edges	0.556	0.525	0.548	0.534	0.506	0.547
# Social Edges	0.538	0.529	0.545	0.530	0.510	0.547
# Conn. Comps, Full Graph	0.579	0.591	0.603	0.525	0.511	0.535
# Conn. Comps, Social Graph	0.567	0.575	0.593	0.527	0.504	0.533
# Conn. Comps, Info. Graph	0.566	0.565	0.585	0.528	0.504	0.546
# Singletons, Full Graph	0.583	0.599	0.602	0.529	0.513	0.548
# Singletons, Social Graph	0.574	0.579	0.595	0.525	0.516	0.548
# Singletons, Info. Graph	0.571	0.566	0.590	0.525	0.513	0.545
Max Comp. Size, Full Graph	0.573	0.587	0.599	0.521	0.520	0.541
Max Comp. Size, Social Graph	0.555	0.581	0.598	0.520	0.520	0.537
Max Comp. Size, Info. Graph	0.556	0.565	0.586	0.527	0.504	0.547
Majority Vote	0.518	0.521	0.547	0.518	0.521	0.547

Table 2: Accuracy of a logistic regression mode for predicting whether a hashtag will double the number of adopters at different starting points: the 1000, 2000, and 4000 initial adopters, for both the follower and the @-message graphs. The accuracy of all models was evaluated using 10-fold cross-validation.

relationship as informational. For example, celebrities on Twitter are followed my many of their fans, but they don't usually follow their fans back. Hence, we could think of Twitter as a network composed in two kinds of edges: social and informational [3]. Given this lack of consistency of the meaning of connections on Twitter, we try to tease apart the two kinds of connections as they may have different prediction potential.

Formally, given the directed follower network on Twitter which we refer to as the *full graph*, we define an undirected edge between users A and B as *informational* if either A follows B, or B follows A, but not both. We define an undirected edge between users A and B as *social* if they follow each other. Next, we define the *social graph* as the network of Twitter users and their social edges only, and the *informational graph* as the users with their informational edges only. Note that the informational and social graphs are undirected, and the full follower graph is directed. Note that we can define corresponding graphs using @-message edges instead of the follower edges in a similar way.

The predictive model. We train a logistic regression model to predict whether the number of adopters of a hashtag will eventually double. We use simple topological properties of the full graph, the social graph, and the informational graph. For each graph we compute the number of edges, number of singletons, number of connected components (weakly for the full follower graph), and the size of the largest connected component. We also train a separate logistic regression model with the same features but based on the @-message graph.

To understand our ability to meaningfully separate graphs based on the structural features we analyze, in Figure 4.1 we plot the complementary cumulative density functions for the features, as computed for all hashtags that exceeded size 1000, 2000, and 4000. We see the number of singletons, the number of components, and the number of adoptees in the largest component, the features are broadly distributed across their support, consistently for all three subgraph sizes.

For each feature, we include its value as well as the logarithm of the value. Additionally, for every feature except for number of edges, we include a "distance from the mid-point" transformation: $|v_f - \frac{m_f}{2}|$, where v_f is the value of the feature and m_f is the largest possible value of the feature. The reason for this transformation is that, as figure 3 suggests, high growth of the hashtags may be correlated with large or small values of some features. Having this

transformation allows the algorithm to capture this trend. We do not include this transformation for number of edges because for these features m_v is extremely large, and none of the hashtags we consider have an edge density greater than 0.5, making all these transformed features linearly dependent upon the initial feature.

We begin by using the logistic regression model to predict whether a hashtag will double its size. For each hashtag h that has at least k adopters, we compute the features of the model and predict whether it will eventually obtained $2k$ adopters. We run the algorithm using the follower graph and the @-message graph separately. For the @-message graph we used a threshold of at least 3 @-messages to form an edge. We did the experiment with different choices of threshold and obtained similar results. Table 4.3 shows the accuracy of the full multivariate model using all the features and transformations as well as a model using each feature and its transformation alone. We evaluated the accuracy using 10-fold cross-validation for $k = 1000, 2000, 4000$. Using the follower graph we obtain an accuracy of around 67%. This is 14 percentage points above a baseline of around 53% obtained from a naive majority vote algorithm, which simply classifies all hashtags as "yes" if the majority of hashtags doubled and "no" if the majority of hashtags do not. Using the @-message graph we do not perform as well as with the follower graph. For the @-graph, the accuracy of the full model about 57% compared to a baseline of about 53%. Furthermore, we find that the accuracy changes very little for the different choices of k, which suggests that classifying hashtags doubling does not get harder of easier we change the original size of the hashtag. Also, we compare the performance with different sets of features, i.e., *Social Graph Only*, *Full Graph Only* or *Informational Graph Only*. It is shown that *Social Graph Only* cannot provide as good performance as the other two feature sets. The reason might be that informational relationships work better in the spread of hashtags. Furthermore, the *Informational Graph Only* performs marginally better than the social graph, and in some case it performs marginally better than the full model will all the features included. It is an interesting open problem to investigate if information edges indeed carry more predictive power than other kinds of edges, and if so, to determine possible explanations for it.

4.4 Predicting Short and Long term Horizons

Having found that the original size of the hashtag does not affect the accuracy of the algorithm, we now investigate whether the accuracy changes as we change the horizon of prediction. That is, what happens to the accuracy if we try to predict whether the hashtag will grow by a factor of p for $p \in (0, \infty)$? We expect that when p is close to 0 the algorithm will not gain much accuracy above the baseline for two reasons. First, the outcome will be very sensitive to noise since we are asking whether the hashtag will obtain just a few additional adopters, so finding the few hashtags that do not surpass the threshold becomes difficult. Second, because most hashtags will surpass the threshold, even the naive majority vote classifier will have high accuracy, leaving little room for improvement. Similarly, when p is very large, we expect that the structure of the graph will lose predictive power as it is itself changing when additional users adopt the hashtag. Also, since most hashtags will not surpass the threshold, the majority vote classifier will have high accuracy, again, leaving little room for improvement.

To answer this question, we run the logistic regression algorithm with the same features as above using the hashtags that had at least 1000 adopters and predicted whether they would reach at least M users. Figure 5 shows the accuracy, precision, recall, and F1 score of the full logistic model as a function of M. We compare these scores with two baseline naive classifiers – The majority vote classifier discussed above, and a random algorithm that classifies as "yes" a random set of hashtags of size equal to the fraction of hashtags that obtained at least M adopters.

We find that, indeed, the accuracy of our classifier is not above the baseline for large and small values of M. However, when we look at the precision of our classifier, we see that it stays above the baselines even for large values of M. That means that even for long term horizon prediction, the structure of the initial graphs maintains predictive power. The recall of the classifier starts off reasonably large for small values of M, but drops as M gets very large. This is due to the fact that when M is large, the number of hashtags that surpass M adopters will be very small, making it very hard to identify all the few that did.

In summary, our classifier maintains an accuracy of roughly 70%. Its accuracy stays above our baselines for mid-term horizons and it equals the baselines for long and short term horizons. Its precision decreases slightly as the horizon increases, but it always stays above our baselines. Its recall starts out high, but it drops dramatically as the horizon increases. It is optimal for a classifier to have high precision and recall, and in our case we are able to maintain high precision, but recall falls for long term horizons. Depending on the actual application of the classifier, sometimes it may be more desirable to have high recall than precision and vice-versa. A possible applications for wanting to know the future size of a certain hashtag is for market forecasting. If people are using hashtags that correspond to new products, and an analyst would like to use social media analysis to track which product to invest in, having high precision is preferable over having good recall. On the other hand, low recall simply means that many products that became popular were not identified by the algorithm, so opportunities were lost, but not investments.

5. RELATED WORK

Collaborative tagging systems, which allow users to share their tags for particular resources, form the basis of our understanding of hashtags [21, 24, 9, 31, 33, 7, 29]. Marlow et al. [21] performed an early study on Flickr, developing a taxonomy of social tagging systems. They found that friends show a larger similarity in vocabulary compared with a random user baseline, which suggested that social links and tag usage is indeed related. Ramage el al. [24] proposed a generative model based on Latent Dirichlet Allocation (LDA) that jointly models text and tags in such settings. Markines et al. [20] employed similarity measures such as matching, overlap, mutual information, and Jaccard, Dice, and cosine similarity to study topics. Recent work by Schifanella et al. [29] also works on the interplay of the social and semantic components of social media on Flicker and Last.fm. They showed that a substantial level of local lexical and topical alignment is observable among users who lie close to each other in the social network. Their analysis suggests that users with similar topical interests are more likely to be friends, and therefore semantic similarity measures among users based solely on their annotation metadata should be predictive of social links.

Models developed to describe small world networks within structural embeddings is closely related to our conceptual approach to hashtags as well. Kleinberg [13] proposed a distance-dependent small world random graph model with a lattice embedding, which was generalized to an hierarchical and a set-based model [14]. Liben-Nowell et al. [23] studied the problem of geographic greedy routing. They empirically observe that the probability of being friends with another person was inversely proportional to the other person's rank distance, the number of people closer to the first individual.

Twitter has been employed as a testbed in many studies due to

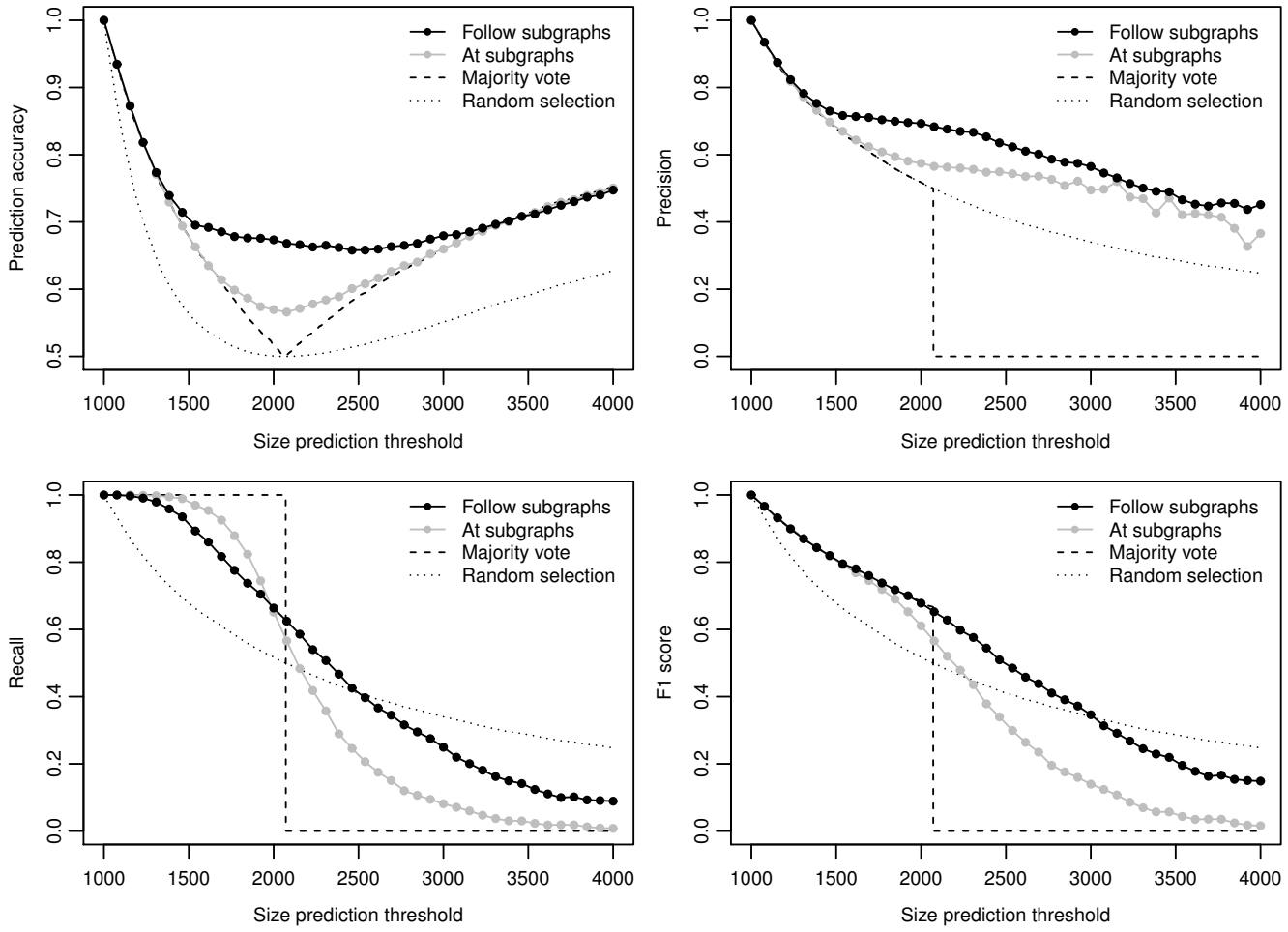

Figure 5: Prediction accuracy, precision, recall, and F1 score when predicting wether a hashtag will exceed a certain size using our logistic regression model based on graph structure. Models were trained using 5-fold cross validation, applied to those 7397 hashtags that reached a size of 1000.

its openness. Kwak et al. [16] presented a myriad of perspectives on Twitter as a microblogging service, and generously made their large "follow" network available. Hashtags have attracted less research, but several recent papers have begun to investigate their role in social media [16, 27, 28, 32, 10]. Romero et al. [27] studied Twitter as an information and social network and the process of directed closure (implicit "link copying") for follow edges, while Romero et al. [28] studied how hashtags diffuse on Twitter. They found significant variation in the ways that widely-used hashtags on different topics spread. Weng et al. [32] attempted to measure the interestingness of hashtags in Twitter.

6. CONCLUSION

In this study, we find that the user-generated hashtag set system on Twitter and the topology of the connections among users are two fundamentally related structures. By studying distance measures among users, based on the topical proximity embedded by the hashtag set system, we are able to predict, with reasonably high accuracy, the links between the users. Furthermore, the size of the smallest common hashtag turns out to be a very good predictor of linkage despite being one of cheapest ones to compute. We also found that combining the topical overlap of the users with the structural features of the graph induced by the shared topics can dramatically improve the accuracy of the prediction task. A possible applications of this would be situations where one knows the topical interests of users and would like to predict connections among them, including recommendations systems for connections in online social networks. For example, recommending who to follow on Twitter based on an individual's hashtag usage, or recommending people to friend on Facebook based on an individual's "likes."

After observing how simple structural features of the graph induced by a hashtag were useful in predicting social connections, we discovered that they are also useful for predicting the popularity of the hashtag itself. These features are very efficient to compute in $O(|V| + |E|)$ time. We found the future popularity of the hashtags does not monotonically increase with the density of the graph induced by its initial set of users. Instead, we find that the popularity of the hashtag is highest when the density is either very low or very high. This is an interesting finding that offers a different perspective to the ideas from "viral marketing" where a small number of connections could be considered a negative property with respect to future growth.

Throughout our study we compare our results generated by the follower graph and the @-message graph. We find interesting distinctions among the two. For example, @-connections are easier to predict, but they turn out to be less informative when predicting the popularity of a hashtag from the connections of early adopters. Also, since the @-graph can be viewed as a weighted graph, we are able compare out results on different levels of edge strength by considering graph with only @-edges with at least k @-messages. We find that stronger ties are easier to predict, but they do not provide better or worse predictive power with regard to future hashtag popularity.

7. ACKNOWLEDGMENTS

We thank Jon Kleinberg for helpful discussions. Also, we thank Brendan Meeder for providing data.

References

[1] L.A. Adamic and E. Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[2] Jacqueline Johnson Brown and Peter H Reingen. Social ties and word-of-mouth referral behavior. *Journal of Consumer Research*, 14(3):350–62, 1987.

[3] Justin Cheng, Daniel Romero, Brendan Meeder, and Jon Kleinberg. Predicting reciprocity in social networks. In *ICSC*, 2011.

[4] D. Easley and J. Kleinberg. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[5] J. Goldenberg, B. Libai, and Muller. Using complex systems analysis to advance marketing theory development. *Academy of Marketing Science Review*, 2001.

[6] Daniel Gruhl, R. Guha, David Liben-Nowell, and Andrew Tomkins. Information diffusion through blogspace. In *WWW*, 2004.

[7] Harry Halpin, Valentin Robu, and Hana Shepherd. The complex dynamics of collaborative tagging. In *WWW*, 2007.

[8] Tony Hammond, Timo Hannay, Ben Lund, and Joanna Scott. Social bookmarking tools (i): A general review. *D-Lib Magazine*, 11, 2005.

[9] Ming-Hung Hsu, Yu-Hui Chang, and Hsin-Hsi Chen. Temporal correlation between social tags and emerging long-term trend detection. In *ICWSM*, 2010.

[10] Jeff Huang, Katherine M. Thornton, and Efthimis N. Efthimiadis. Conversational tagging in twitter. In *HT*, 2010.

[11] Matthew O. Jackson. *Social and Economic Networks*. Princeton University Press, August 2008.

[12] Masahiro Kimura and Kazumi Saito. Tractable models for information diffusion in social networks. In *PKDD*, 2006.

[13] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *STOC*, 2000.

[14] J. Kleinberg. Small-world phenomena and the dynamics of information. In *NIPS*, 2002.

[15] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. In *KDD*, 2006.

[16] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is Twitter, a social network or a news media? In *WWW*, 2010.

[17] Jure Leskovec, Lada A. Adamic, and Bernardo A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 2007.

[18] Jure Leskovec, Kevin J. Lang, Anirban Dasgupta, and Michael W. Mahoney. Statistical properties of community structure in large social and information networks. In *WWW*, 2008.

[19] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, 58:1019–1031, May 2007.

[20] Benjamin Markines, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho, and Stum Gerd. Evaluating similarity measures for emergent semantics of social tagging. In *WWW*, 2009.

[21] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. Ht06, tagging paper, taxonomy, flickr, academic article, to read. In *HT*, 2006.

[22] M. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[23] Liben D. Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *PNAS*, 102(33):11623–11628, 2005.

[24] Daniel Ramage, Paul Heymann, Christopher D. Manning, and Hector Garcia-Molina. Clustering the tagged web. In *WSDM*, 2009.

[25] Matthew Richardson and Pedro Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD*, 2002.

[26] Everett M. Rogers. *Diffusion of Innovations, Fourth Edition*. Free Press, 1995.

[27] Daniel M. Romero and Jon Kleinberg. The Directed Closure Process in Hybrid Social-Information Networks, with an Analysis of Link Formation on Twitter. In *ICWSM*, 2010.

[28] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. Differences in the Mechanics of Information Diffusion Across Topics: Idioms, Political Hashtags, and Complex Contagion on Twitter. In *WWW*, 2011.

[29] Rossano Schifanella, Alain Barrat, Ciro Cattuto, Benjamin Markines, and Filippo Menczer. Folks in folksonomies: social link prediction from shared metadata. In *WSDM*, 2010.

[30] Ben Taskar, Ming fai Wong, Pieter Abbeel, and Daphne Koller. Link prediction in relational data. In *NIPS*, 2003.

[31] Jennifer Thom-Santelli, Michael J. Muller, and David R. Millen. Social tagging roles: publishers, evangelists, leaders. In *CHI*, 2008.

[32] Jianshu Weng, Ee-Peng Lim, Qi He, and Cane Wing-Ki Leung. What do people want in microblogs? measuring interestingness of hashtags in twitter. In *ICDM*, 2010.

[33] Zhijun Yin, Rui Li, Qiaozhu Mei, and Jiawei Han. Exploring social tagging graph for web object classification. In *KDD*, 2009.